Task 1: Semi-supervised Learning - Self/Co-Training

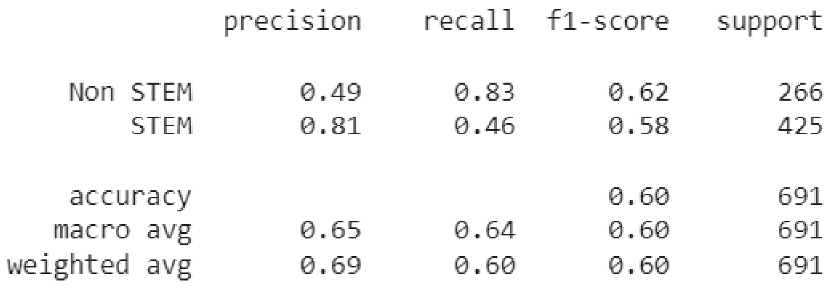
The task was to use Wikipedia API to extract the summaries of random Wikipedia articles and label a part of them as either STEM or NON-STEM subjects. About 30 - 40 labels. The remaining would be unlabelled.

A semi-supervised learning model was to be made. It basically works by training a model on the labelled set and using that model to predict on the remaining dataset. The model also has a confidence threshold for example 70%. Such that it will only add the labels to the remaining set if it is at least 70% sure that it's prediction is correct. This process is repeated a certain number of times until we have a fully trained model.

I used Sklearn library which has a method to train self learning models. It takes a base estimator model and uses that to turn it into a self learning model.

Note that it requires the unlabelled part of the dataset to be set to -1.

Using the data extractor tool, I also created a testing dataset to evaluate the model performance. Here is the output of it:



```
              precision    recall  f1-score   support

    Non STEM       0.49      0.83      0.62       266
        STEM       0.81      0.46      0.58       425

    accuracy                           0.60       691
   macro avg       0.65      0.64      0.60       691
weighted avg       0.69      0.60      0.60       691
```

The model performed decently for STEM subjects but did terrible for Non STEM subjects.

Changing the threshold also didn’t seem to give any better results. The best threshold for label prediction was at 75% confidence.